Protein Engineering vol.13 no.7 pp.459–475, 2000

Identifying sequence–structure pairs undetected by sequence alignments

Sanzo Miyazawa[1,2] and Robert L.Jernigan[3]

[1]Faculty of Technology, Gunma University, Kiryu, Gunma 376, Japan and [3]Room B-116, Bldg 12B, MSC 5677, Laboratory of Experimental and Computational Biology, DBS, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892-5677, USA

[2]To whom correspondence should be addressed.
E-mail: miyazawa@smlab.sci.gunma-u.ac.jp

We examine how effectively simple potential functions previously developed can identify compatibilities between sequences and structures of proteins for database searches. The potential function consists of pairwise contact energies, repulsive packing potentials of residues for overly dense arrangement and short-range potentials for secondary structures, all of which were estimated from statistical preferences observed in known protein structures. Each potential energy term was modified to represent compatibilities between sequences and structures for globular proteins. Pairwise contact interactions in a sequence–structure alignment are evaluated in a mean field approximation on the basis of probabilities of site pairs to be aligned. Gap penalties are assumed to be proportional to the number of contacts at each residue position, and as a result gaps will be more frequently placed on protein surfaces than in cores. In addition to minimum energy alignments, we use probability alignments made by successively aligning site pairs in order by pairwise alignment probabilities. The results show that the present energy function and alignment method can detect well both folds compatible with a given sequence and, inversely, sequences compatible with a given fold, and yield mostly similar alignments for these two types of sequence and structure pairs. Probability alignments consisting of most reliable site pairs only can yield extremely small root mean square deviations, and including less reliable pairs increases the deviations. Also, it is observed that secondary structure potentials are usefully complementary to yield improved alignments with this method. Remarkably, by this method some individual sequence–structure pairs are detected having only 5–20% sequence identity.

Keywords: empirical potentials/inverse protein folding/protein fold recognition/sequence–structure alignment/threading and inverse threading with gaps and insertions

Introduction

Measuring compatibilities between sequences and structures is neither simple nor easy. Hendlich *et al*. (1990) used a set of potentials of mean force from Sippl (1990) to approximate residue–residue interactions, and demonstrated that native structures have lower values than alternative structures, with some understandable exceptions. In their calculations, non-native structures were generated by threading sequences into the structures of other proteins at all possible positions without gaps. The same method to generate non-native folds for sequences has been used by Sippl and Weitckus (1992) and Bryant and Lawrence (1993). Such a method to generate alternative folds in which any gap is disallowed is appropriate, because conformations of a sequence can be compared on the same scale of conformational energy. A simple comparison of conformational energy values between different sequences is meaningless. However, forbidding gaps is too extreme a restraint for generating alternative folds. Park and Levitt (1996) and Park *et al*. (1997) generated many decoys by relaxing native structures with molecular dynamics or other methods, and tested many types of empirical energy functions for their abilities to distinguish correct from incorrect folds. Vendruscolo and Domany (1998) proposed a Monte Carlo dynamics in contact map space to generate uncorrelated low-energy states to serve as decoys.

In order to allow gaps in sequence–structure alignments, two types of problems must be overcome. If deletions and insertions in sequence–structure alignments are to be allowed, then the problem of fold recognition becomes essentially the same as for the inverse protein folding problem (defined as the problem of selecting from a set of sequences only those sequences that are compatible with a single given structure). One must take into account not only the conformational energies of folds but also the sequence dependences of the whole ensemble of protein conformations in order to evaluate the relative stabilities of sequences or alignments (Miyazawa and Jernigan, 1999c). Here, the stabilities of structures are assumed as a primary requirement for compatibilities between sequences and structures. The second problem is how to evaluate multi-body interactions among residues, or at least specifically the pairwise interactions.

From the viewpoint of the inverse folding problem, Bowie *et al*. (1991) developed a method in which the fitness of each type of residue at a given residue position in a structure is evaluated with respect to the environment of the residue in the native structure, and then a conventional dynamic programming method (Needleman and Wunsch, 1970) is utilized to align a sequence with a given structure. The score of the alignment obtained is used to represent the compatibility of the sequence with the given structure. It has also been used to evaluate protein models (Lüthy *et al*., 1992). This method is based on the fact that the environment of a particular residue in a structure is more conservative than the residue itself, and is equivalent to an approximation called the 'frozen approximation' by Godzik *et al*. (1992), in which the residue's environment is evaluated for the native sequence rather than the trial sequence. If the 'frozen approximation' is used, a conventional dynamic programming method can be used for sequence–structure alignment. However, in principle, the assumption of the native structure environment is inappropriate for detecting structural similarities between extremely divergent proteins and especially between proteins sharing a common fold through convergent evolution, where the environments surrounding equivalent residue position could be dissimilar (Jones and Thornton, 1993).

Nishikawa and Matsuo (1993) developed an improved evaluation function by adding hydration potentials, hydrogen bond potentials and local conformational potentials, all of which were estimated as potentials of mean force based on statistical preferences observed in known protein structures. They reported that homologous sequence pairs in a sequence database could also be discriminated on the basis of structure–sequence compatibility. In their work, sequences were aligned on the basis of sequence information only by a conventional dynamic programming method, and then 3D–1D compatibilities of protein pairs were evaluated, although 3D–1D alignments were made with the 'frozen approximation' in their later work (Matsuo *et al*., 1995).

In order to evaluate more precisely pairwise interactions between residues, Jones *et al*. (1992) used a double dynamic programming method that was originally devised for structural alignments by Taylor and Orengo (1989) and which is an approximate method to take account of pairwise potentials. A search algorithm for finding exact global optimum threadings into protein core segments connected by variable loops, was devised for pairwise interaction potentials (Lathrop and Smith, 1996).

These and a number of other works (Crippen, 1991; Finkelstein and Reva, 1991; Maiorov and Crippen, 1992; Sippl, 1993; Kocher *et al*., 1994; Matsuo and Nishikawa, 1994; Huang *et al*., 1995; Park and Levitt, 1996; Thomas and Dill, 1996; Park *et al*., 1997) indicate that simple empirical potentials without atomic details may be sufficient to determine overall folds, although some limitation to pairwise potentials is indicated (Mirny and Shaknovich, 1996). Munson and Singh (1997) developed multi-body potentials for recognition between sequences and structures. Samudrala and Moult (1998) illustrated the importance of using a detailed atomic description for obtaining the most accurate discrimination. To increase weak signals in each pairwise sequence–structure alignment, multiple sequence threading was also utilized (Taylor, 1997).

Here, we examine the utility of the simple potential function developed by Miyazawa and Jernigan (1999c) for identifying compatibilities between sequences and structures of proteins. The potential function consists of pairwise contact energies (Miyazawa and Jernigan, 1985, 1996, 1999a), repulsive packing potentials for residues (Miyazawa and Jernigan, 1996) and short-range potentials for secondary structures (Miyazawa and Jernigan, 1999b). These potentials were estimated from statistical preferences observed in known protein structures, but are devised to represent the actual interactions in proteins and to be able to estimate actual conformational energies for a wide range of conformations from the native to the denatured state.

Miyazawa and Jernigan (1999c) described how to modify these energy potentials to represent approximately the stabilities of proteins, both multimeric and monomeric, and also how to define a single reference state that is appropriate for both fold and sequence recognition. There, it was shown that this simple scoring function can distinguish native structures from alternate folds and also discriminate native sequences from non-native sequences, in which non-native sequence and structure pairs are generated by threading sequences into other structures in all possible ways without gaps. Here, it will be generalized by allowing deletions and insertions in sequence–structure alignments. Gap penalties will be assumed to be proportional to the number of contacts at each residue position so that gaps tend to be more frequently placed on protein surfaces than in cores.

We evaluate pairwise contact interactions between residues in a mean field approximation on the basis of the probabilities of site pairs being aligned. To obtain the self-consistent values of alignment probabilities of site pairs, an iterative method is employed. In addition to the most probable alignment, that is, the minimum energy alignment, an alignment is also made by successively assigning aligned site pairs by their alignment probabilities; see Miyazawa (1995) for this alignment method. This method, termed a probability alignment, can provide information regarding how reliable each individual aligned pair is. This feature is certainly desirable for aligning distantly related sequences and structures. No information about native amino acid sequences but only structural information is used in the present sequence–structure alignments, in order to see how well empirical energy potentials can select for compatibilities between sequences and structures.

Materials and methods

Conformational energy

The total conformational energy of a protein is represented here as a sum over contributions from residues along the sequence as

$$E^{\mathrm{conf}} \equiv \sum_{p} E_p^{\mathrm{conf}} \tag{1}$$

Each residue's contribution is further divided into two terms, for secondary structure and for tertiary structure:

$$E_p^{\mathrm{conf}} \equiv E_p^{\mathrm{sec}} + E_p^{\mathrm{tert}} \tag{2}$$

where p indexes residue position.

The short-range interaction energies for secondary structures used here are those estimated (Miyazawa and Jernigan, 1999b) by a potential of mean force from the observed frequencies of secondary structures in known protein structures, which are assumed to be in an equilibrium distribution following the Boltzmann factors of their secondary structure energies. The effects of long-range interactions are taken into account only as a mean field. Because of the limited number of available protein structures, the secondary structure potential is approximated as a sum of additive contributions from neighboring residues along a sequence, with neglect of side chain–side chain interactions. Non-additive contributions are simply neglected. In addition, the effects here from neighboring residues are limited to a dependence on their amino acid types but not on their secondary structures. The conformational specification is limited to a sequential tripeptide. Thus, their secondary structure potential is approximated as a sum of the following contributions:

$$E_p^{\mathrm{sec}} \approx e^s(s_{p-1},s_p,s_{p+1}) + \sum_{p-3\le q\le p+3} \delta e^s(s_{q-1},s_q,s_{q+1},i_p) \tag{3}$$

where i_p is the pth residue of type i_p and s_p is the conformational state of the pth residue. The first term in Equation 3 represents the backbone–backbone interactions and the second term corresponds to side chain–backbone interactions either within a residue or among close residues. Altogether side chain–backbone interactions within five consecutive backbone units on each side of a side chain are included in the short-range interactions. Here it should be noted that two-body and higher order interactions between side chains and backbones of triplets are counted only once in the estimation of each term in

Table I. Average energies per residue expected for randomly aligned residues

Category	Energy per residue in kT units				
	Of the native structure (mean)	For random alignments			
		Mean[a]	S.d.	Mean[b]	S.d.
Secondary structure energies	0.0	0.83	1.37		
Alignment contact energies	0.0	0.80[a]	1.80	1.14[b]	1.67
Repulsive energies	0.0	–0.06[a,c]		–0.08[b,c]	
Total energies	0.0	1.57[a]		1.89[b]	

[a]The average energy of randomly aligning a residue in the native structural environment.
[b]The average energy per residue in threading randomly shuffled sequences into structures.
[c]These are negative values because they are a subtraction of excess contact energies due to tight packing.

Equation 3 to add to the total short-range interaction. The first term $e^s(s_{-1},s_0,s_1)$ is also defined (Miyazawa and Jernigan, 1999b) to include only half of the two-body interactions between nearest neighbors in order to avoid multiple counts of nearest neighbor interactions in the estimation of the total secondary structure energy of Equation 1.

The tertiary structure energies have previously been estimated as a sum of pairwise residue–residue contact energies and repulsive residue packing energies for volume exclusion, together termed long-range interaction energies in Miyazawa and Jernigan (1996):

$$E_p^{\text{tert}} = E_p^{\text{c}} + E_p^{\text{r}} \tag{4}$$

The contact energy E_p^{c} and the repulsive packing energy E_p^{r} of a residue at position p are defined by Equations 18, 19 and 40 in one of our previous papers (Miyazawa and Jernigan, 1996). For the contact energies (e_{ij}) for all pairs of the 20 types of residues, which are applied to residue–residue close contacts, our estimates (Miyazawa and Jernigan, 1999a) corrected for the Bethe approximation are used here. Actually, the new estimates of contact energies listed in Miyazawa and Jernigan (1999a) are divided by $\alpha' \approx 0.263$ defined in Equation 34 in that paper and used as the values of contact energies. In other words, the intrinsic pairwise interaction energies (δe_{ij}) are corrected relative to the hydrophobic energies (Δe_{ir}), and the hydrophobic energies are not corrected at all, in order to make the magnitude of contact energies comparable to secondary structure energies (see Table I). The repulsive packing energies for the 20 types of residues corresponding to penalties for overly dense packing, which are a function of the number of residues in contact, previously estimated by us (Miyazawa and Jernigan, 1996), are employed here.

The contact energy E_p^{c} and repulsive packing energy E_p^{r} of residues at each position in structures as required in Equation 4 are calculated according to Equations 18, 19 and 40–43 in Miyazawa and Jernigan (1996). However, the hard core repulsion term is not included here, i.e. e^{hc} is set to zero in their Equation 41, since there should not be such extremely overly dense regions in any properly refined structures.

Alignment energy for scoring of sequence–structure compatibility

The stability of native structure is assumed as a primary requirement for proteins to fold into their native structures. The probability $\mathcal{P}(\{s_p\}|\{i_p\})$ with which a protein sequence $\{i_p\}$ takes a specific conformation $\{s_p\}$ is represented by its conformational energy relative to the free energy of the whole ensemble of protein conformations:

$$-\log[\mathcal{P}(\{s_p\}|\{i_p\})] = \beta E^{\text{conf}}(\{s_p\}|\{i_p\}) + \log\left[\sum_{\{S_p\}} \exp(-\beta E^{\text{conf}}(\{s_p\}|\{i_p\}))\right] \tag{5}$$

$$\approx \beta \Delta E^{\text{conf}}(\{s_p\}|\{i_p\}) + n_r\sigma \tag{6}$$

where

$$\Delta E^{\text{conf}}(\{s_p\}|\{i_p\}) \equiv E^{\text{conf}}(\{s_p\}|\{i_p\}) - (E^{\text{conf}} \text{ of a typical native structure with the same amino acid composition}) \tag{7}$$

$$\sigma \equiv \text{conformational entropy per residue in } k \text{ units for native-like structures} \tag{8}$$

β is equal to $1/(kT)$ and n_{r} is the sequence length of a protein. In Equation 5, $E^{\text{conf}}(\{s_p\}|\{i_p\})$ is the conformational energy of state $\{s_p\}$ of sequence $\{i_p\}$, and the sum is taken over all possible conformations. Therefore, the free energy of the whole ensemble can be regarded as a zero energy state, i.e. a reference state for the energy potential representing protein stability. The free energy of the protein ensemble varies unless the protein sequence is fixed. Thus, in order to discuss the compatibilities of different protein sequences with a given fold, the change of the protein ensemble due to variable sequence must be taken into account, in addition to the conformational energy. In sequence–structure alignments, deletions in amino acid sequences are allowed, so that the change to the whole ensemble of protein conformations must be taken into account. As discussed in detail in Miyazawa and Jernigan (1999c), the second term of Equation 5 is approximated by Equation 6; in the sum of Boltzmann factors over all conformations only dominant terms, i.e. native-like compact conformations are taken into account, and then the logarithm of the function is evaluated in a high-temperature expansion. σ, representing a conformational entropy per residue for native-like structures, is a constant independent of the amino acid sequence of the protein. The unweighted average of $E^{\text{conf}}(\{s_p\}|\{i_p\})$ over native-like conformations is approximated as the conformational energy expected for a typical native structure with the given amino acid composition, which depends only on amino acid composition. This approximation is justified by testing of sequence recognition in inverse threading without gaps performed in our previous work (Miyazawa and Jernigan, 1999c).

If the environments surrounding proteins are the same, the stabilities of those proteins can be compared by potential energies with proper reference states. However, in fold and sequence recognition protein conformations in a monomeric state may need to be compared with other structures in a multimeric state. In this case, entropy loss due to binding ought to be taken into account in addition to binding energies between subunits, in order to measure the stabilities of protein structures in a multimeric state. To overcome this difficulty, energy potentials are modified to measure approximately protein stabilities even for proteins in different environments. A collapse energy (e_{rr}) is subtracted from the contact energies to remove the protein size dependences and in order to represent protein stabilities for monomeric and multimeric states. On the other hand, the intrinsic potential and backbone–backbone interaction potentials for secondary structures depend strongly on the types of protein structures from which they are estimated. Thus, only energy terms dependent on amino

acid sequences are included in a scoring function for sequence–structure alignments; the first term in Equation 3 is ignored and only the other terms are included. These modifications for energy potentials used here have been discussed in detail in Miyazawa and Jernigan (1999c).

After all of the considerations above are included, the following quantity is taken for assessing compatibilities between protein sequences and structures:

$$\Delta E_p^{\rm conf}\,(e_{i_pj}-e_{\rm rr}) \equiv \Delta E_p^{\rm sec} + \Delta E_p^{\rm tert}(e_{i_pj}-e_{\rm rr}) \qquad (9)$$

where $e_{ij} - e_{\rm rr}$ within parentheses means that it is the argument of the function. Because judgements on insertions and deletions in sequence–structure alignments are made for every residue, these modifications to energies are taken into account for every residue.

The energy score for secondary structures is defined as follows, by excluding the intrinsic and backbone–backbone interaction energies:

$$\Delta E_p^{\rm sec} \approx \Delta e^s(...,s_{p-1},i_p,s_p,s_{p+1},...) \equiv \sum_{p-3\leqslant q\leqslant p+3} \delta e^s(i_p,s_{q-1},s_q,s_{q+1}) \qquad (10)$$

The reference state for the backbone–side chain interaction potentials $\delta e_s(i_p,s_{q-1},s_q,s_{q+1})$, is defined by adjusting their average energies over native structures to zero energy.

For the tertiary structure energies, the reference energy corresponds to the average tertiary structure energy per residue for each type of residue in the native protein structures. That is, the following difference in the tertiary structure energy is considered:

$$\Delta E_p^{\rm tert} \equiv \Delta E_p^{\rm c} + \Delta E_p^{\rm r} \qquad (11)$$

$$\equiv (E_p^{\rm c} - <E_{i_p}^{c}>) + (E_p^{\rm r} - <E_{i_p}^{\rm r}>) \qquad (12)$$

The second and the fourth terms in Equation 12 are the average contact energy per residue of type i_p and the average repulsive energy per residue of type i_p in native structures.

Alignment by evaluating pairwise interactions in a mean field approximation

An example of a specific sequence–structure alignment A is

$$A \equiv \begin{bmatrix} ... - i_3\, i_4\, i_5\, i_6 ... \\ ... \quad s_2 - s_3 - s_4 ... \end{bmatrix} \qquad (13)$$

where – means a deletion, s_p is the conformational state of the pth residue in a given structure and i_q is the qth residue of type i_q in sequence that is threaded into the structure.

From Equation 6, a conditional probability $\mathcal{P}(\{s_p\}|\{i_p\},A)$ with which the sequence in alignment A takes a specific conformation $\{s_p\}$ can be approximated as follows:

$$-\log\{\mathcal{P}(\{s_p\}|\{i_q\},A)\} \approx \beta \sum_{(p,q)\in A} \Delta E_p^{\rm conf}\,(\{s_p\}|i_q,A) + n_r^{\rm aligned}\sigma \qquad (14)$$

where $n_r^{\rm aligned}$ is the number of aligned site pairs in the alignment A. $\Delta E_p^{\rm conf}\,(\{s_p\}|i_q,A)$ is an alignment energy for aligning the qth residue i_q, of the sequence to the pth residue structure position, s_p, in the structure $\{s_p\}$ threaded with the aligned energy A.

Then, according to Bayes' rule (Feller, 1968), the conditional probability $\mathcal{P}(A|\{s_p\},\{i_q\})$ of an alignment A for a given structure $\{s_p\}$ is represented as

$$\mathcal{P}(A|\{s_p\},\{i_q\}) = \mathcal{P}\,(\{s_p\}|\{i_q\},A)\;\mathcal{P}(A) \Big/ \Big[\sum_A \mathcal{P}(\{s_p\}|\{i_q\},A)\mathcal{P}(A)\Big] \qquad (15)$$

where $\mathcal{P}(A)$ is the *a priori* probability for an alignment A. Here, this *a priori* probability is represented as follows by introducing penalties for gaps:

$$-\log\{\mathcal{P}(A)\} \equiv n_r^{\rm aligned}(\beta\mathcal{E}_0 - \sigma) + \beta\Big[\sum_{\rm all\ gaps\ in\ } \mathcal{W}\Big] + {\rm constant} \qquad (16)$$

where $\mathcal{W}$ is a positive quantity to represent a gap penalty and $\mathcal{E}_0$ is a negative constant that does not depend on the correspondence of the qth residue i_q to the residue position p and is used as a scaling parameter together with gap penalties; it is chosen in such a way that the total energy scores of sequence–structure alignments for random sequences are always positive. In the case of gapless alignments, i.e. simple threadings of sequences into a fold, the *a priori* probability is the same for any alignment, because the second term in Equation 16 is equal to zero. Unless alignments between different sequence and structure pairs are compared, the use of $\mathcal{E}_0$ is equivalent to the addition of $-\mathcal{E}_0/2$ to the gap penalty in one scheme for gap penalty, because $k_{\rm i} + k_{\rm d} + 2k_{\rm m} = n_{\rm r}^{\rm seq} + n_{\rm r}^{\rm str}$, where $k_{\rm i}$, $k_{\rm d}$ and $k_{\rm m}$ are the numbers of insertions, deletions and matches/mismatches, respectively, and $n_{\rm r}^{\rm seq}$ and $n_{\rm r}^{\rm str}$ are the lengths of the sequence and structure, respectively. However, not all gaps are equivalent in the present scheme where different gap penalties are employed for terminal gaps and for the middle of gaps, and also the gap penalty is taken to have an upper limit; the parameter choices are described later.

Thus, the conditional probability of an alignment A for a given structure $\{s_{\rm p}\}$ is represented as

$$\mathcal{P}(A|\{s_p\},\{i_q\}) = \frac{1}{\mathcal{Z}} \exp[-\beta\mathcal{E}(\{s_p\}|\{i_q\},A)] \qquad (17)$$

$$\mathcal{Z} = \sum_A \exp[-\beta\mathcal{E}(\{s_p\}|\{i_q\},A)] \qquad (18)$$

where $\mathcal{Z}$ is a partition function for alignments. The energy score $\mathcal{E}\,(\{s_p\}|\{i_q\},A)$ of an alignment A for a given structure $\{s_p\}$ is defined as

$$\mathcal{E}(\{s_p\}|\{i_q\},A) \equiv \sum_{(p,q)\in A} \mathcal{E}\,(\{s_p\}|i_q,A) + \sum_{\rm all\ gaps\ in\ } \mathcal{W} \qquad (19)$$

The energy score $\mathcal{E}(\{s_p\}|i_q,A)$ is simply equal to the alignment energy $\Delta E_p^{\rm conf}$ with a scaling parameter

$$\mathcal{E}(\{s_p\}|i_q,A) \equiv \Delta E_p^{\rm conf}\,(\{s_p\}|i_q,A) + \mathcal{E}_0 \qquad (20)$$

Pairwise interactions are evaluated on the basis of the probabilities for site pairs to be aligned, that is, this is a kind of mean field approximation. Thus, the pairwise interaction energies in $\Delta E_p^{\rm conf}\,(\{s_p\}|i_q,A)$ for alignment A are approximated with pairwise energies for amino acid pairs (i_q,i'_q) located at neighboring sites (p,p') in structure with alignment probabilities $\mathcal{P}\,(p',q')$ of structure–sequence site pairs (p',q'):

$$\mathcal{E}(\{s_p\}|i_q,A) \approx \mathcal{E}(\{s_p\}|i_q,\mathcal{P}(p',q')) \qquad (21)$$

The probability for a structure–sequence site pair (p,q) to be aligned and the probabilities for deletions $(p,–)$ and $(–,q)$ are calculated as

$$\mathcal{P}(p,q) = \frac{1}{\mathcal{Z}} \sum_{A\ {\rm with}\ (p,q)} \exp[-\beta\mathcal{E}(\{s_p\}|\{i_q\},A)] \qquad (22)$$

$$\approx \frac{1}{\mathcal{Z}} \mathcal{Z}_{p-1,q-1} \exp[-\beta \mathcal{E}(\{s_p\}|i_q,\mathcal{P}(p',q'))] \mathcal{Z}'_{p+1,q+1} \quad (23)$$

$$\mathcal{P}(p,-) = 1 - \sum_q \mathcal{P}(p,q) \quad (24)$$

$$\mathcal{P}(-,q) = 1 - \sum_p \mathcal{P}(p,q) \quad (25)$$

where $\mathcal{Z}_{p-1,q-1}$ is also a partition function but for aligning the N-terminal, partial sequence from 1 to (q–1)th residues with the N-terminal, partial structure from 1 to (p–1)th residues in the whole structure. $\mathcal{Z}'_{p+1,q+1}$ is a partition function for aligning the C-terminal sequence starting from the (q+1)th residue with the C-terminal part from p+1 to the terminal end in the whole structure. Therefore, the following relation is satisfied:

$$\mathcal{Z} = \mathcal{Z}_{n_r^{str}, n_r^{seq}} = \mathcal{Z}'_{1,1} \quad (26)$$

where n_r^{seq} and n_r^{str} are the lengths of the sequence and structure, respectively. Such partition functions can be calculated from energy scores by a transfer matrix method; see Miyazawa (1995) for a specific description of this method for alignments. To obtain a self-consistent solution for alignment probabilities $\mathcal{P}(p,q)$ of structure–sequence site pairs (p,q) in Equation 23, an iteration method is employed here.

$\Delta E_p^{conf}(\{s_p\}|i_q,\mathcal{P}(p',q'))$ in Equations 20–21 is calculated as the sum of contributions of short-range, secondary structure interactions and long-range, tertiary structure interactions; see Equation 9:

$$\Delta E_p^{conf}(\{s_p\}|i_q,\mathcal{P}(p',q')) \equiv \Delta E_p^{sec}(\{s_p\}|i_q,\mathcal{P}(p',q')) + \Delta E_p^{tert}(\{s_p\}|i_q,\mathcal{P}(p',q')) \quad (27)$$

The present evaluation of the secondary structure energies does not include side chain–side chain interactions, so that the secondary structure energy, the first term in the above equation, is calculated as the short-range interaction energy between the backbone conformation (...,s_p,...) and the qth residue of type i_q placed at the pth residue position of the structure. In other words, the secondary structure energies have only single body interactions with respect to side chains; hence they can be evaluated without requiring alignment information.

$$\Delta E_p^{sec}(\{s_p\}|i_q,\mathcal{P}(p',q')) \approx \Delta e^s(...,s_{p-1},i_q,s_p,s_{p+1},...) \quad (28)$$

where Δe^s is defined by Equation 10. The long-range component, $\Delta E_p^{tert}(\{s_p\}|i_q,\mathcal{P}(p',q'))$, which is defined by Equations 9, 11 and 12, and which includes pairwise contact energies between residues and density-dependent packing interactions among residues, is evaluated as an alignment energy for aligning the amino acid i_q at the pth residue position s_p of the target structure on the basis of the alignment probabilities for site pairs obtained in the previous iteration. For the first iteration, Godzik *et al*. (1992) employed the native sequence to evaluate the environment surrounding a residue; instead, the alignment that does not have gaps except at both termini and in which residue i_q is forced to be aligned to position s_p is assumed here for evaluating the long-range energies.

Then, by evaluating the energy score of alignments with the self-consistent alignment probabilities of site pairs in Equation 21, we can easily calculate the minimum energy score alignment that is the most probable alignment with a conventional dynamic programming method; the minimum energy score alignment A^{min} is defined as

$$\mathcal{E}(\{s_p\}|\{i_q\},A^{min}) = \min_A \mathcal{E}(\{s_p\}|\{i_q\},A) \quad (29)$$

$$\approx \min_A \left[\sum_{(p,q)\in A} \mathcal{E}(\{s_p\}|i_q,\mathcal{P}(p',q')) + \sum_{\text{all gaps in } A} \mathcal{W} \right] \quad (30)$$

The approximation in Equation 21 and this approximation in Equation 30 for the minimum energy alignment becomes rigorous in the low-temperature limit.

In addition, we also employ here alignments consisting of the most probable site pairs by successively aligning a site pair in order of pairwise alignment probabilities $\mathcal{P}(p,q)$ as follows (Miyazawa, 1995):

(i) Set p_1 and p_2 to the N-terminal and C-terminal site position of a sequence segment to align, and q_1 and q_2 to the N-terminal and C-terminal site position of a partial structure to align.

(ii) If there is a site pair (p,q) such that $\mathcal{P}(p,q) = \max_{p_1\le p'\le p_2, q_1\le q'\le q_2}(\mathcal{P}(p',q')|\mathcal{P}(p',q') \ge \mathcal{P}(p',-)$ and $\mathcal{P}(p',q') \ge \mathcal{P}(-,q'))$, align them. Otherwise, assign deletions to all sites of $p_1 \le p' \le p_2$ and of $q_1 \le q' \le q_2$. Then, repeat steps (i) and (ii) to align the remaining segments until all the sites are aligned.

Here we term such an alignment a probability alignment. It should be noted that a probability alignment is different from the most probable alignment, i.e. the minimum energy score alignment. The former is based on alignment probabilities of site pairs, and the latter simply means the alignment with the maximum probability, that is, the minimum energy score. Of course, the probability alignment coincides with the most probable alignment in the limit of low temperature.

A whole ensemble of sequence-structure alignments can be characterized by such quantities as the minimum energy score, free energy score, and internal energy score. The minimum energy score and the free energy score are defined as:

$$\mathcal{E}_{min} \equiv \min_A \mathcal{E}(\{s_p\}|\{i_q\},A) \quad (31)$$

$$\mathcal{F} \equiv -\frac{1}{\beta} \log \mathcal{Z} \quad (32)$$

The statistical average of energy scores over all alignments, which corresponds to the internal energy, is calculated from the following relation involving the partition function:

$$<\mathcal{E}(\{s_p\}|\{i_q\},_A)>_A = -\frac{1}{\mathcal{Z}} \frac{\partial \mathcal{Z}}{\partial \beta} \quad (33)$$

A preliminary test indicates that the capability of recognition of sequence–structure compatibilities seems to be about the same among these three energy scales. In the following, minimum energy scores are employed to judge sequence–structure compatibilities.

Gap penalty for sequence–structure alignments

The effects of amino acid replacements on protein structure are not uniform over a sequence, indicating a dependence of amino acid variabilities on residue position. It is well known that, on average, residues are more conserved in the interiors of proteins than on their surfaces (Go and Miyazawa, 1980). One may expect deletions and insertions of residues to occur

more frequently in less conserved regions of a sequence. Gap penalties ought to reflect the mutability at each residue position. Here the dependence of residue mutability on residue position (Go and Miyazawa, 1980) is taken into account by setting the gap penalty to be proportional to the number of contacts at each residue position in a protein structure. The number of contacts is utilized here as a simple measure of burial and packing density of residues; see Equations 34–36. In other words, gaps will tend to be inserted in alignments more often on protein surfaces than in protein cores.

For deletions of residues of the structure in sequence–structure alignments, a gap penalty is taken as

$$\mathcal{W} = \min\left(w_0 + \sum_{p \in \text{gap}} (w_1 + w_2 n_p^{\text{c}}), w_{\text{c}}\right) \tag{34}$$

where w_0, w_1, w_2 and w_{c} are parameters taking zero or positive values and n_p^{c} is the number of residues in contact with the pth residue. The definition of a contact is the same as that for contact energies; if the two centers of side chains are within 6.5 Å, then they are defined to be in contact with each other. The summation in the equation above is taken over all deletions in a gap. The value of the gap penalty is cut off beyond a certain value w_{c} to allow us to find single domains in multi-domain proteins, and also to reduce computational time for alignments.

For insertions of k residues to the sequence in sequence–structure alignments, inserted between the qth and $(q+1)$th residue positions in the structure, the gap penalty is set to

$$\mathcal{W} = \min(w_0 + k(w_1 + w_2(n_q^{\text{c}} + n_{q+1}^{\text{c}})/2), w_{\text{c}}) \tag{35}$$

Equation 35 is trivially different for additions of terminal residues

$$\mathcal{W} = \min(w'_0 + k(w'_1 + w'_2 n_{\text{terminal}}^{\text{c}}), w'_{\text{c}}) \tag{36}$$

where w'_0, w'_1, w'_2 and w'_{c} are all parameters taking zero or positive values.

Here, it should be noted that the gap penalties are convex functions of gap length because all the gap parameters above take zero or positive values. Gap parameters could be set to different values for insertions and for deletions; however, the same values are employed here for both, in order to reduce the number of parameters. On the other hand, in general, penalties for terminal gaps and gaps in the middle of the sequence should have different values. The algorithm of maximum similarity alignment (minimum energy alignment) used for the present gap scheme corresponds to Equations 22 and 23 in Miyazawa (1995).

Parameter choice for sequence–structure alignments

In conventional sequence alignments, it is well known that an alignment program can produce significantly different alignments with different parameter settings. The effects that a parameter choice has on resulting alignments have been studied (Fitch and Smith, 1983; Vingron and Waterman, 1994). Gotoh (1990) also studied the effects of the variation of gap penalties. Heuristic knowledge about gap penalties in conventional sequence alignments is used in sequence–structure alignments.

Parameters which we must specify are deletion penalties ($\mathcal{E}_0$ for aligning residues, w_0, w_1, w_2, w_{c} for gaps in the middle, and w'_0, w'_1, w'_2, w'_{c} for terminal gaps) and the relative temperature $1/\beta$. First, let us consider one of the parameters, $\mathcal{E}_0$, which is an additional energy for aligning a residue at a structural position against a deletion and an insertion, and is a scaling parameter independent of the residue type and residue position; see Equations 16 and 19–20 for its definition. The parameter $\mathcal{E}_0$ is chosen in such a way that minimum energy scores for most of the dissimilar protein pairs falls above zero; also there is no clear indication that the minimum energy scores depend linearly on the sequence length. If this were not the case, long or short alignments would tend to have low-energy scores independent of whether proteins aligned were related.

Table II. Gap parameters used in sequence–structure alignments[a]

Gap penalty	Value in kT units
$\mathcal{E}_0$	–1.2
Structure deletions from q to q_1	$5.5 + \sum_{p=q}^{q_1} (1.05 + 0.43 n_p^c)$ in the middle[b] $3.25 + \sum_{p=q}^{q_1} (0.53 + 0.22 n_p^c)$ at termini
n sequence insertions between q and $q+1$	$5.5 + n\,[1.05 + 0.43\{1 + (n_q^c + n_{q+1}^c)/2\}]$ in the middle $3.25 + n\,[0.53 + 0.22(1 + n_{terminal}^c)]$ at termini
The upper limits for gap penalty	60.9 for gaps in the middle 30.45 for terminal gaps
Relative temperature, $1/\beta$	2.6

[a]These parameter values are for a case in which secondary structure energies, contact energies and repulsive energies are all included.
[b] n_p^c is the number of residues in contact with the pth residue.

The mean energy for random alignments of residues is listed for each type of interaction potential in Table I. All energies in the following are represented in kT units. Here we emphasize that the mean of each energy component of short-range secondary structure energy, contact energy and repulsive energy, in native structures is set to zero; see Equations 9–12. Permitting gaps in alignments improves energy scores over the mean energy scores for random residue matches. Thus, $\mathcal{E}_0$ must be more positive than –1.57 [= –(0.83+0.80–0.06)] with secondary structure energies included or –0.74 [= –(0.80–0.06)] without secondary structure energies.

The penalties for a deletion or an insertion of a residue must be greater than one half of $-\mathcal{E}(\{s_p\}|i_q, A)$, that is the score for aligning the qth residue of type i_q at position p as defined by Equation 20, because the sum of sequence lengths of the two proteins is equal to the sum of the numbers of deletions, insertions, and two times the number of other matches or mismatches in an alignment. Otherwise alignments would not be favored. In the present case, the largest average individual increment of tertiary structure alignment energy in the native environment is expected to be 3.39 (4.13 for contact energies) for misaligning a Leu to a Lys position and 3.67 (4.03 for contact energies) for misaligning a Cys to a Lys position in the native environment, and for secondary structure energies to be 5.36 for misaligning Pro to a Gly position [see Table III of Miyazawa and Jernigan (1999b)]. The largest average increments of tertiary structure alignment energies are smaller for misalignments in the random environment than for those in the native structure environments. The largest increment of tertiary structure alignment energy, 4.97 (5.33 for contact energies), will occur if Leu is aligned to residue positions that are completely exposed to water. On the other

hand, the largest average increment of the sums of secondary structure energies and tertiary structure energies is 5.66 (6.19 for contact energies) for misaligning Ile to a Gly position in the native environment, and 5.90 (6.44 for contact energies) for the random environment. Based on such information, the parameters defined by Equations 34–36 for gap penalties are given in Table II. w_0+w_1 has been set to be greater than $(5.90+\mathcal{E}_0)/2$ with secondary structure energies or $(4.97+\mathcal{E}_0)/2$ without secondary structure energies. w_1+4w_2 has been configured to be greater than $(5.90+\mathcal{E}_0)/2$ with secondary structure energies or $(3.67+\mathcal{E}_0)/2$ without secondary structure energies; the average number of residues in contact at each residue position in proteins is 4.19. The value of the gap penalty is cut off beyond a certain value w_c to avoid loading too much penalty onto long gaps. The use of upper limits for gap penalties is especially appropriate for global alignments, over a whole sequence, of sequence–structure pairs in which a compatible domain is limited to only a portion of the sequence and structure. The value for an upper limit w_c is chosen to be equal to a penalty for a gap of 20 residues on average; $w_c \sim w_0+20(w_1+4w_2)$. Based on choices in conventional sequence alignments (Miyazawa, 1995), the gap penalties and their upper limits have been set to be smaller for terminal gaps than for gaps in the middle of a sequence. The parameters for terminal gaps, w'_0, w'_1, w'_2 and w'_c, are arbitrarily set to be one-half of their corresponding parameter values for middle gaps, w_0, w_1, w_2 and w_c, respectively. All gap parameters used here are listed in Table II.

As shown in the Results section, the present values of gap parameters are adjusted to yield similar fractions of aligned residues in minimum energy alignments for homologous protein pairs to those in conventional sequence alignments. The relative temperature (1/β) is also adjusted to yield similar fractions of aligned residues in probability alignments for the homologous protein pairs compared with those in probability sequence alignments.

Datasets of protein structures

Two datasets of protein pairs were prepared; one is a set of homologous protein pairs, and the other is a set of dissimilar protein pairs. For each protein pair in these sets, we calculate minimum energy alignments and also probability alignments, and examine whether their sequences and structures are compatible with each other.

Release 1.35 of the SCOP database (Murzin *et al*., 1995) is used for the classification of protein folds. Representatives of superfamilies, families or domains are the first entries in the protein lists of each superfamily, each family or each domain in SCOP; if these first proteins in the lists are not appropriate to use for the present purpose, then the second ones are chosen. These families and domains are all those which belong to the protein classes 1–5, that is, classes of all α, all β, α/β, α+β, and multi-domain proteins. Classes of membrane and cell surface proteins, small proteins, peptides and designed proteins are not used. Proteins whose structures were determined by NMR or with resolution worse than 2.5 Å are removed. Also, proteins whose coordinate sets either consist of only C^α atoms, or include many unknown residues, or lack many atoms or residues, are removed. Proteins shorter than 50 residues are also removed.

In the SCOP database, protein domains whose sequences are highly homologous may be classified into the same domains, and protein domains whose structures are extremely similar may belong to different domains although in the same family. Therefore, protein pairs, which are more similar than 90% sequence identity, or whose structures are more similar than 1 Å r.m.s.d. (root mean square deviation), are also removed from the set of domain representatives. As a result, our set of superfamily representatives includes 308 proteins, the set of family representatives has 440 proteins and the set of domain representatives has 988 proteins.

The set of homologous protein pairs is made by pairing the protein representatives of families with those of different domains within the families; the number of homologous protein pairs in this set is 548. Because there are families that consist of only one domain present, only 164 families are included in this set. The set of dissimilar protein pairs is made by arbitrarily choosing only every 100th pair from the ordered list of all possible pairs of superfamily representatives; 505 protein pairs are chosen. In sequence–structure alignments, the first proteins in those protein pairs are used as sequences and the second ones as structures; in other words, the sequences of family representatives and the structures of domain representatives in the same families are compared in sequence–structure alignments of homologous proteins. In inverse structure–sequence alignments, the first proteins are used as structures and the second ones as sequences.

Table III. Parameters used in the conventional maximum sequence similarity alignment

Parameter	Value
Amino acid similarity score	Dayhoff's 250-PAM log-odd matrix[a] divided by[b] ln(2)/3
Gap penalty for *n* insertions/deletions	$12 + 4(n-1)$ in the middle of a sequence $6 + 2(n-1)$ at termini
Upper limits of gap penalties	48 in the middle of a sequence 24 at termini
Temperature-like parameter[c]	3/ln(2)

[a]Dayhoff *et al*. (1978).
[b]Karlin and Altschul (1990).
[c]See Miyazawa (1995) for details.

Results

Adequacy of sequence–structure alignments

First, the adequacy of sequence–structure alignments with the present method has been examined by comparing the overall characteristics of sequence–structure alignments with those of conventional sequence alignments. Both secondary structure energies and tertiary structure energies are included in the calculation of alignment energy scores. Folds of multimeric proteins and domains are evaluated in the multimeric state or within a whole protein even for sequences of monomeric proteins. Table III shows the values of gap parameters used here for our conventional sequence alignment; the Dayhoff 250 PAM matrix (Dayhoff *et al*., 1978) is used as a scoring matrix for the sequence alignment, but alternatively BLOSUM matrices (Henikoff and Henikoff, 1992) could have been used.

Figure 1 shows comparisons of the fractions of aligned residue pairs and the fractions of identical amino acid pairs for 548 homologous protein pairs between sequence–structure alignments and conventional sequence alignments. In Figure 1a and 1c, minimum energy score alignments defined by Equation 30 are compared with maximum similarity alignments for sequences. In Figure 1b, probability alignments, which are

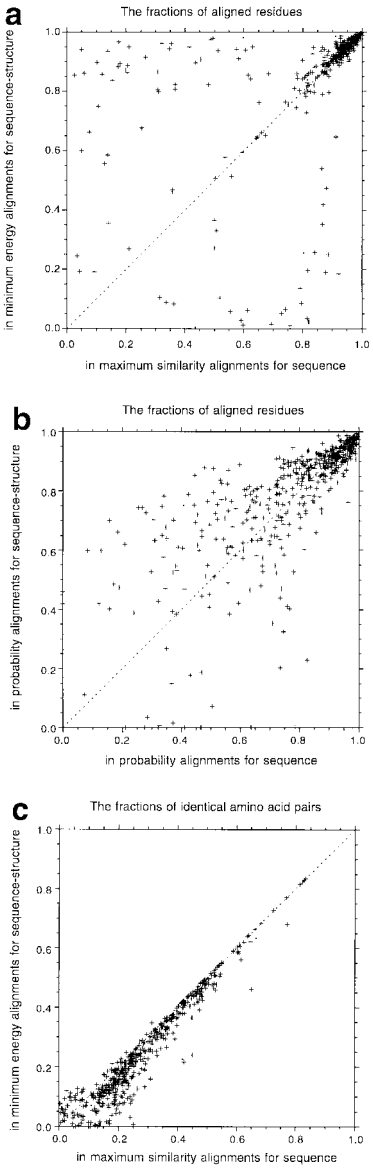

Fig. 1. Comparisons of overall characteristics between sequence–structure alignments and conventional sequence alignments for homologous protein pairs. Minimum energy score alignments for sequence–structure are compared with maximum similarity alignments for sequence regarding the fraction of aligned residues in (**a**) and also regarding the fraction of identical amino acid pairs in (**c**). Probability alignments for sequence–structure are compared with those for sequence regarding the fraction of aligned residues in (**b**). Both secondary structure energies and tertiary structure energies are included in the calculation of alignment energy scores. The set of 548 homologous protein pairs consisting of family representatives versus domain representatives within the same families from the SCOP database is employed here. In the sequence–structure alignments above, family representatives are used as sequences and domain representatives are employed as structures. The dotted lines show lines with equal values for both axes.

made by successively aligning site pairs in order of their alignment probabilities, are employed for both sequence–structure and sequence–sequence alignments. The fraction of aligned residue pairs is defined as

$$\text{Fraction of aligned residue pairs} \equiv \frac{2\,(\text{number of aligned residue pairs in the alignment})}{(\text{length of sequence 1})+(\text{length of sequence 2})} \quad (37)$$

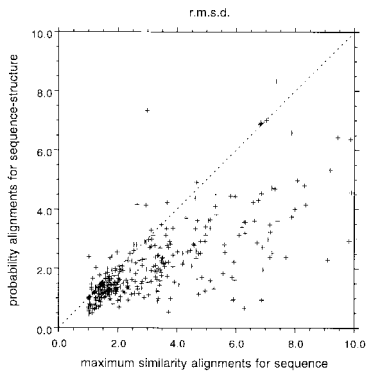

Fig. 2. Comparisons of the r.m.s.d.s in superposition of aligned residue pairs in probability alignments of sequence–structure with those in maximum similarity alignments of sequences. Only residue pairs aligned with probability $\geqslant$0.5 are used for sequence–structure alignments. 357 homologous protein pairs, where both types of alignments have at least 50 aligned residue pairs, whose minimum energy scores are more negative than zero and maximum similarity scores are more positive than zero, are plotted in this figure; only protein pairs with r.m.s.d. $<$10 Å are shown here. The dotted line shows a line with equal values for both axes.

The fraction of identical amino acid pairs is defined in a similar way.

Both the sequence–structure alignments and the conventional sequence alignments give similar aligned fractions of residues for most proteins, indicating the values of $\mathcal{E}_0$ and gap parameters to be appropriate. Also, as shown in Figure 1b, we have adjusted the relative temperature (1/β) in such a way that similar fractions of residues are aligned in the probability alignments for both sequence–structure and sequence–sequence alignments.

As shown in Figure 1c, the present method of sequence–structure alignments yields only slightly fewer identical amino acid pairs than the conventional sequence alignment method, especially for relatively dissimilar proteins. This is understandable, since the sequence–structure alignment method does not actually maximize the sequence identity, as does the conventional sequence alignment method.

Figure 1 indicates the adequacy of sequence–structure alignments for homologous protein pairs for their overall characteristics. To examine further the quality of the present sequence–structure alignments, the r.m.s.d.s in superpositions of aligned residue pairs in the sequence–structure alignments are compared in Figure 2 with those from the maximum similarity alignments of sequences. For this purpose, we employ probability alignments for sequence–structure consisting of only the most reliable residue pairs aligned with probabilities $\geqslant$0.5. The r.m.s.d.s of aligned residue pairs calculated for dissimilar protein pairs indicate that values of r.m.s.d. can be $<$7 Å even for dissimilar protein pairs, if the number of superposed residues is $<$50. Therefore, in this figure, protein pairs whose alignments have $<$50 aligned residue pairs are excluded. In addition, homologous protein pairs which have positive minimum energy scores and negative maximum similarity scores are excluded: in other words, only homologous protein pairs whose similarities are identified by both methods are used. The 357 homologous protein pairs meeting these criteria are plotted in this figure for the sequence–structure alignments. Significant improvements in the values of r.m.s.d. are shown. Although these improvements are made partially by choosing only residue pairs most reliably aligned, they also indicate that the quality of the probability alignments

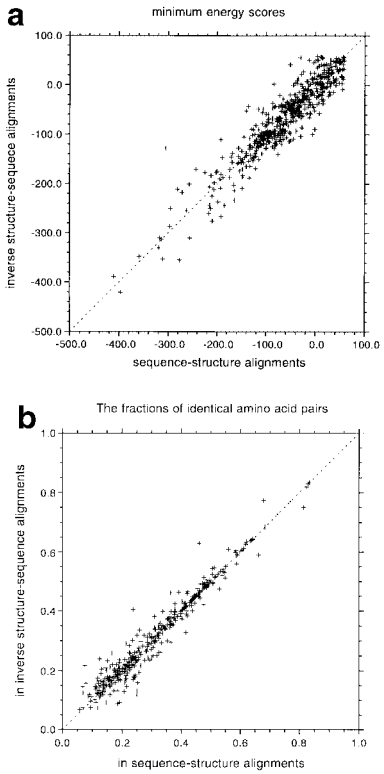

Fig. 3. Comparisons between sequence–structure alignments and the inverse structure–sequence alignments for the same homologous protein pairs. The minimum energy scores of alignments, and the fractions of identical amino acid pairs in the alignments are shown in (**a**) and (**b**). The set of 548 homologous protein pairs is employed here. In the sequence–structure alignments above, family representatives are used as sequences and domain representatives are employed as structures, and in the inverse structure–sequence alignments family representatives are used as structures and domain representatives as sequences. To clearly see correlations between the two types of alignments, only 391 protein pairs whose homology is detected to be significant, i.e. where both the types of alignments have more negative scores of minimum energy than zero for the present parameter values, are plotted in (b); in each alignment of those protein pairs, $\geqslant$50 residue pairs are aligned.

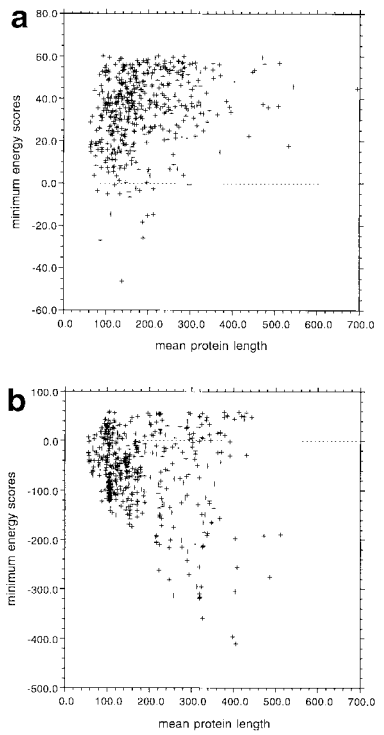

Fig. 4. The distribution of minimum energy scores for sequence–structure alignments over the means of the lengths of each protein pair in (**a**) for dissimilar protein pairs and in (**b**) for homologous protein pairs. The set of 548 homologous protein pairs, and the set of 505 dissimilar protein pairs between superfamily representatives are employed here. The dotted lines are for zero energy score.

of sequence–structure are usually better than those for the corresponding conventional sequence alignments.

Comparison of two types of sequence–structure alignments

Two different types of sequence–structure alignments can be utilized to assess the similarity between two proteins from the viewpoint of sequence–structure relationships: using the first as sequence and the second as structure and then inversely using the first as structure and the second as sequence. Figure 3 shows comparisons between sequence–structure alignments and their inverse structure–sequence alignments for the same homologous protein pairs. In the sequence–structure alignments, family representatives are used as sequences and domain representatives are employed as structures, and in the inverse structure–sequence alignments, structures are family representatives and sequences are domain representatives.

Minimum energy scores for the alignments are compared in Figure 3a. Both types of alignment give similar minimum energy scores. Overall characteristics such as the fractions of aligned residue pairs, the fractions of identical amino acid pairs in the alignments, and r.m.s.d.s in superposition of aligned residue pairs have been compared between the two types of alignments; only the fractions of identical amino acid pairs are shown in Figure 3b. To see clearly correlations between both types of alignment, only the 391 protein pairs whose similarities are detected to be significant, i.e. whose minimum energy scores are below zero for the present parameter values, are compared. As expected, both types of alignment take similar values for the fraction of aligned residues, for the fraction of identical amino acid pairs and for the r.m.s.d.s of aligned residues, although slightly different alignments may be obtained with the two types of alignments.

One interesting observation is that on average the energy scores for the alignments are roughly equal for the two types of alignments; see Figure 3a. This result indicates that the present scale of energies and its reference state may be used equally well either to detect compatible sequences with a given structure or compatible folds for a given sequence.

Detection of homologous proteins from dissimilar proteins

One of the most important questions is how well this energy scale can recognize a compatible pair of structure and sequence, particularly those not found from sequence comparisons. The minimum energy scores of alignments are plotted in Figure 4a for the dissimilar protein pairs and in Figure 4b for the homologous protein pairs; see the section ''Datasets of protein structures'' above for these protein datasets.

The parameter $\mathcal{E}_0$ is chosen so that in Figure 4a there is no clear indication that the minimum energy scores of the dissimilar structure pairs depend linearly on the lengths of proteins and also so that most of those minimum energy scores fall above zero; otherwise, long or short alignments would tend to have low energy scores independently of whether proteins aligned are related. Because gap penalties are cut off beyond the values in Table II, the total minimum energy scores of sequence–structure alignments are limited by an upper bound,

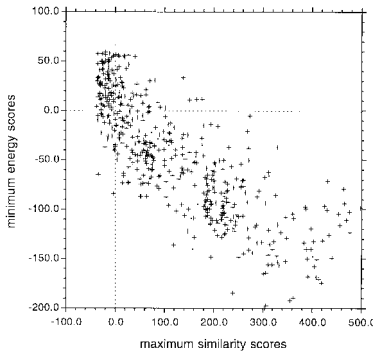

Fig. 5. Comparison of the minimum energy scores of sequence–structure alignments with the maximum similarity scores of conventional sequence alignments for homologous protein pairs. This figure shows only protein pairs whose maximum similarity scores are <500. The dotted lines are lines with zero scores for each axis. The numbers of protein pairs in the regions of negative-positive, negative-negative positive-positive, and positive-negative values of abscissa and ordinate are 81, 25, 50 and 392, respectively. The correlation coefficient is 0.85.

the sum of the cut-off values of gap penalties at both termini, which is equal to 60.9 in the present case; see Table II. The average of energy scores per residue for native sequence–structure pairs is equal to the value of $\mathcal{E}_0$, –1.2; this value is almost a lower bound for minimum energy scores per residue.

As shown in Figure 1c, the present set of homologous protein pairs includes many distantly related protein pairs whose alignments have fractions of identical amino acid pairs below 10%. Thus, as shown in Figure 4b, there are many distantly related protein pairs which have positive minimum energy scores of alignment and are not identified as compatible sequence–structure pairs. The conventional sequence alignment method cannot detect similarities for all of those homologous protein pairs, either. Table IV lists the numbers of false positives and false negatives for the present sequence–structure alignment method and for the conventional sequence alignment method. Here, the judgements are made solely on the basis of the values of scores. In sequence–structure alignments, gap parameters are adjusted so that compatible sequence and structure pairs tend to take negative energy scores and incompatible ones positive energy scores. However, in conventional sequence alignments, gap parameters are adjusted so that positive scores are expected for similar sequences and negative scores for dissimilar sequences. The overall capability to identify homologous protein pairs is slightly better for the conventional sequence method than for the present sequence–structure alignment method, but both methods can complement each other to recognize some different homologous protein pairs. Figure 5 shows a comparison of alignment scores between both methods. On the basis of the values of scores, both methods identify similarities for 392 protein pairs of the 548 homologous protein pairs, but fail for 81 protein pairs. The present sequence–structure alignments identify 25 homologous protein pairs whose similarities were not identified by the conventional sequence alignment method. In the case of the inverse structure–sequence pairs, both methods identify the similarities of 395 protein pairs but fail for 79 protein pairs. The inverse structure–sequence alignments can identify the similarities of 27 homologous protein pairs that cannot be identified by the sequence alignments; 11 of those 27 protein pairs are protein pairs whose compatibilities are identified in common by both the sequence–structure alignments and inverse structure–sequence alignments.

Table IV. Recognition of homologous protein pairs from dissimilar protein pairs by sequence–structure alignments

False positives in 548 homologous protein pairs	False negatives in 505 dissimilar protein pairs	Method
106	5	Conventional sequence alignment
131	17	Sequence–structure alignment
126	24	Inverse structure–sequence alignment
191	19	Sequence–structure alignment without secondary structure energies
164	26	Inverse structure–sequence alignment without secondary structure energies

To establish that those alignments are reasonable, the r.m.s.d.s of the sequence–structure alignments are examined. To ensure that the r.m.s.d. are reliable, only protein pairs having ⩾50 residue pairs aligned with probabilities ⩾0.5 are listed in Table V. The relatively small r.m.s.d. values for these protein pairs in sequence–structure alignments indicate that reasonable alignments for most of the protein pairs are obtained. One of the most interesting protein pairs in this list is the pair UDP–galactose 4-epimerase from *Escherichia coli* (1XEL) and human estrogenic 17-β-hydroxysteroid dehydrogenase (1FDS) whose alignment includes less than 10% sequence identity.

To judge whether such alignment scores are statistically significant, one may use a *z*-score that is defined as an alignment score expressed in standard deviation (s.d.) units from the average score for randomized sequences. Figure 6 shows that the present energy scores roughly correlate with the *z*-scores evaluated from 100 randomized sequences, and that a zero energy score corresponds to about –3 s.d. units; the correlation coefficient is 0.81.

It is also useful to know the relationship between minimum energy scores of alignments and similarities between structures. In Figure 7a, minimum energy scores per residue are plotted against r.m.s.d.s in the superposition of residues aligned with probabilities ⩾0.5. To reduce the effects of the number of aligned residues on the value of r.m.s.d., homologous protein pairs with aligned residue pairs ⩾50 are plotted in this figure. Most of the probability alignments whose minimum energy scores fall below zero energy score, which is a threshold for identifying compatible proteins, have r.m.s.d.s <5 Å. Especially, if the minimum energy scores per residue are more negative than –0.2, then almost all alignments with few exceptions have r.m.s.d.s <5 Å. Interesting cases appear if one looks closely at the exceptional protein pairs.

Four protein pairs with r.m.s.d. >10 Å and with minimum energy scores per residue more negative than –0.5 are troponin C from chicken (1NCX) sequence compared with the 1TCO-B, 1WDC-C, 1WDC-B and 1LIN structures in the calmodulin-like family. There is a helix in the middle of the sequences whose lengths vary among these proteins. Thus, even though the structures of both terminal domains are similar, the r.m.s.d. takes on large values. Also, four protein pairs having r.m.s.d.s between 6 and 7 Å and minimum energy scores per residue more negative than –0.8 include 1NCX; 1NCX sequence aligned with other calmodulin-like structures 1CLL, 3CLN, 1OSA and 4CLN structures.

The alignment of the immunoglobulin chain A from Fab HIL (8FAB-A:3–105) and the CD2 chain A from rat (1CDC-

Table V. Protein pairs[a] whose compatibilities are not identified by sequence alignments but by sequence–structure alignments

Sequence	Length	Structure	Length	Sequence–structure				Sequence–sequence			
				Minimum energy alignment		Probability alignment		Maximum similarity alignment			
				Minimum energy score	Fraction of identical residues	No. of residues aligned with probabilities ⩾0.5	R.m.s.d.	Maximum similarity score	Fraction of identical residues	No. of aligned residue pairs	R.m.s.d.
1ECF-A:250–469	220	1HMP-A	214	–10.7	0.11	88	4.6	–11	0.14	193	15.3
1NCX	162	2SAS	185	–17.3	0.10	85	9.1	–6	0.14	161	14.5
1PBN	289	1ECP-A	237	–6.5	0.09	99	5.4	–25	0.02	27	8.0
1PII:1–254	254	1TTQ-A	256	–12.3	0.12	62	11.8	–22	0.03	36	9.2
1PTV-A	297	1YTS	278	–36.2	0.11	105	4.9	0	0.19	260	9.5
1XEL	338	1ENY	268	–3.1	0.05	57	10.9	–2	0.12	189	18.2
1XEL	338	1FDS	282	–20.2	0.10	61	2.6	–1	0.05	54	13.7
2DRI	271	2LBP	346	–26.4	0.13	157	7.3	–14	0.15	211	23.1
2DRI	271	2LIV	344	–37.1	0.11	165	8.1	–20	0.04	63	17.2
2HVM	273	1NAR	289	–84.2	0.11	103	4.0	–3	0.17	266	6.1
2HVM	273	2EBN	285	–22.7	0.10	111	10.1	–28	0.04	59	8.3
2OHX-A:175–324	150	1QOR-A:136–265	130	–40.2	0.21	99	4.9	–1	0.22	127	6.0
3GRS:364–478	115	1NPX:322–447	126	–26.4	0.12	73	3.0	–6	0.13	115	17.1
8FAB-A:3–105	103	1HNF:4–104	101	–39.3	0.11	61	2.8	–2	0.12	98	3.9

[a]Only protein pairs having ⩾50 residue pairs aligned with probabilities ⩾0.5 are listed in this table.

Fig. 6. Comparison of minimum energy scores and their *z*-scores, defined as scores in standard deviation units from the average scores for randomized sequences. In the set of 548 homologous protein pairs, only protein pairs whose minimum energy scores are greater than –200 are shown in this figure. The means and standard deviations of scores for randomized sequences have been estimated from 100 shuffled sequences. The dotted lines show lines with zero energy score and with –3 s.d. units. The solid line is the regression line, $y = -2.8 + 0.08x$; the correlation coefficient is 0.81.

Fig. 7. The relationships (**a**) between minimum energy scores per residue and r.m.s.d. in superposition of residues aligned with probabilities ⩾0.5 in sequence–structure alignments and (**b**) between minimum energy scores per residue and the fractions of identical amino acid pairs in minimum energy score alignments. A minimum energy score per residue is defined as a minimum energy score divided by the mean length of a protein sequence and structure. The set of 548 homologous protein pairs is employed here. In (a), only 398 protein pairs with aligned residue pairs ⩾50 are plotted. The dotted lines are for zero energy score.

A) has a negative minimum energy score per residue of –0.45, but has an extremely large r.m.s.d. value, 19.7 Å; this protein pair is not shown in Figure 7a, because the number of residue pairs aligned with probabilities ⩾0.5 is only 47, which is <50. The reason for the large r.m.s.d. is that the coordinates of 1CDC correspond to a metastable misfolded structure. It is interesting that such a misfolded structure has been detected

to be compatible with a sequence for which the alignment contains a fraction of identical amino acid pairs below 0.15.

Figure 7b shows the relationship between minimum energy scores per residue and the fractions of identical amino acid pairs in the minimum energy score alignments for the 548 homologous protein pairs. This figure indicates that almost all protein pairs having fractions of identical amino acid pairs >0.2 have negative minimum energy scores, and thus can be identified to be similar. Remarkably, some protein pairs with fractions of identical amino acid pairs <0.10 can have negative energy scores and can therefore be identified to be compatible. The strength of the new approach presented here lies in the individual cases newly identified to be similar, which are not found by sequence comparisons.

Examples of sequence–structure alignment

Figure 8a shows sequence–structure alignments between human glutathione reductase C-terminal domain of residues

a

```
minimum energy alignment
    sequence  3GRS 364    YNNIPTVV-FSHPPIGTVGLTEDEAIHKYGIENVKTYSTS FTPMYHAVTKRKTKCVM
    matched to:                        |  |  |  | |    |                  | |
    structure 1NPX 322    GVQGSSGLAVFDYKFASTGINEVMA-QKLGK-ETKAVTVV -EDYLMDFNPDKQKAWF
probability alignment
    sequence  3GRS 364    YNNIPTVV-FSHPPIGTVGLTEDEAIHKYGIENVKTYSTS FTPMYHAVTKRKTKCVM
    matched to:                  ?     |  |  ?  | | |  |   ?              | |
    structure 1NPX 322    GVQGSSGLAVFDYKFASTGINEVM-AQKLGKE-TKAVT-V VEDYLMDFNPDKQKAWF
                          77776643343346989998875415777764243332O3 34444444455566666

              1NPX 322        bbbbb  bbbbbb  aaaa aaaa     bbbb b bbbb        bbbbb
                ######                ######                          #########
              3GRS 364          bbb     bbbbbbbaaaaaaaaaa  bbbbbbbb b aaaaa        bbb
minimum energy alignment
    structure 3GRS 364    YNNIPTVVFSHP PIGTVGLTEDEAIHKYGIENVKTYSTS FTPMYHAVTKRKTKCVM
    matched to:                              | | |                 |        | |
    sequence  1NPX 322    GVQGSSGLAVFD YKFASTGINEVMAQKLGKETKAVTVVE DYLMDF--NPDKQKAWF
probability alignment
    structure 3GRS 364  --YNNIPTVVFSHP-PIGTVGLTEDEAIHKYGIENVKTYSTS-FTPMYHAVTKRKTKCVM
    matched to:           ??      ?     |  |?  | | |     ?? ???      | |
    sequence  1NPX 322  GV--QGSSGLAVFDYKFASTGINE-VMAQKLGKETKAVTVVEDY---LMDFNPDKQKAWF
                        432233444444430345556554145667776543322222021112233566677777


minimum energy alignment
    sequence  3GRS 420  KMVCANKEEKVVGIHMQG-LGCDEMLQGFAVAVKMGATKADFDNT-VAIHPTSSEE   L
    matched to:          | |          |                  |      |   |          |
    structure 1NPX 376  KLVYDPETTQILGAQLMSKADLTANINAISLAIQAKMTIEDLAYADFFFQPAFDKP   W
probability alignment
    sequence  3GRS 420  KMVCANKEEKVVGIHM-QGLGCDEMLQGFAVAVKMGATKADFDNT-VAIHPTS-SEE- L
    matched to:          | |          |   ?              |      |   |    ?     |      ?
    structure 1NPX 376  KLVYDPETTQILGAQLMSKADLTANINAISLAIQAKMTIEDLAYADFFFQPAFDKPWN I
                        666666677777765404567999999888888888888887643445543322221 2

              1NPX 376  bbbb        bbbbbb       aaaaaaaaaa  aaaaaaa             a a
                        #####             ##########################
              3GRS 420  bbbbb b  bbbbbbbbbb      aaaaaaaaaa  aaaaaaa           aaaaa
minimum energy alignment
    structure 3GRS 420  KMVCA-NKEEKVVGIHMQGLGCDEMLQGFAVAVKMGATKADFDNT----VAIHPTSSEEL
    matched to:          | |          |                  |      |   |
    sequence  1NPX 376  KLVYDPETTQILGAQLMSKADLTANINAISLAIQAKMTIEDLAYADFFFQPAFDKPWNII
probability alignment
    structure 3GRS 420  KMVCANKEEKVVG-IHMQGLGCDEMLQGFAVAVKMGATKADFDN----TVAIHPTSSEEL
    matched to:          | |          |?  |                  |      |   |       ?     ??????
    sequence  1NPX 376  KLVYDPETTQILGAQLMSKADLTANINAISLAIQAKMTIEDLAYADFF------------
                        776653443444303444334444445555567778888887643356222222111100


minimum energy alignment                                    min.ene. rmsd #aligned identities
    sequence  3GRS 475  VTLR              -----------
    matched to:                                              -26.4    3.9  112      0.12
    structure 1NPX 433  NIIN              TAALEAVKQER
probability alignment
    sequence  3GRS 475  VTLR              ------------
    matched to:          ??                                            3.7  108      0.12
    structure 1NPX 435  I---              NTAALEAVKQER
                        2011              246789999999                  3.0   73

              1NPX 435  a                 aaaaaaaaaaa
                                              ########
              3GRS 475  aa
minimum energy
    structure 3GRS 475  VTLR                  --------
    matched to:          |                                   -20.0    4.3  113      0.11
    sequence  1NPX 436  NTAA                  LEAVKQER
probability alignment
    structure 3GRS 475  VTLR-------------------------
    matched to:         ???                                           3.5   92      0.12
    sequence  1NPX 424  ----FQPAFDKPWNIINTAALEAVKQER
                        011253334323334434457778899 9                 3.0   45
```

Fig. 8. *continued*

364–478 (3GRS:364–478) and NADH peroxidase C-terminal domain of residues 322–447 from *Enterococcus faecalis* (1NPX:322–447). Both types of alignment, that is, the sequence of 3GRS:364–478 versus the structure of 1NPX:322–447, and inversely the structure of 3GRS:364–478 versus the sequence of 1NPX:322–447, are shown. Also, for each type of sequence–structure alignment, two kinds of alignment are shown in this figure: the minimum energy score alignment and the probability alignment that is made by successively aligning site pairs in order of their alignment probabilities.

This protein pair is one of the protein pairs whose compatibility was not detected by the conventional sequence alignment, but only by the present sequence–structure alignment. As shown in Table V, the maximum similarity score is negative, and the r.m.s.d. of the maximum similarity alignment is >17 Å. On the other hand, as shown at the end of each alignment in Figure 8a, both types of sequence–structure alignment have negative minimum energy scores and r.m.s.d. <4 Å. The fractions of identical amino acid pairs in the sequence–structure alignments are ⩽0.12.

Superposition of the two structures, 3GRS (residues 364–478) and 1NPX (residues 322–447), with the 73 matched residues, which have a probability of ⩾0.5 of being aligned in the probability alignment of the sequence of 3GRS with the structure of 1NPX, is shown in Figure 8b. These aligned residues are shown in green for 3GRS and in blue for 1NPX. It can be seen that the green and blue regions constitute a type of core of these structure fragments.

The minimum energy alignments and probability alignments tend to align the same residue pairs but not always, when alignment probabilities are >0.5; also, it should be noted that both types of sequence–structure and inverse structure–sequence alignments tend to be identical especially at sites aligned with probabilities >0.5; sites commonly aligned in all alignments are marked by '#' between the alignments. This fact indicates the suitability of the present scoring function for both fold and sequence recognition.

Figure 9 shows the sequence–structure alignments of purine nucleoside phosphorylase from bovine (1PBN) and purine nucleoside phosphorylase A chain from *E.coli* (1ECP-A). In this figure, probability alignments show only those residues aligned with probabilities ⩾0.5. This protein pair is also one of the protein pairs whose compatibilities were not detected by the conventional sequence alignment, but only by the present sequence–structure alignment. There are at least two sequence alignments with the maximum similarity score (–25), which are completely different from each other, for this protein pair. One alignment yields only a small number of aligned residues, i.e. 27 residue pairs, and the other aligns as many as 231 residue pairs; the fraction of identical amino acids is 0.02 for the former and 0.14 for the latter. The r.m.s.d. of 27 residue pairs for the former is 8.0 Å, which is attained due to such a small number of superposed residues. The r.m.s.d. for the latter is 15.4 Å. These facts indicate the present sequence alignment method actually fails to find similarities for this protein pair. On the other hand, the r.m.s.d. for the minimum energy score alignment of the 1PBN structure with 1ECP-A sequence is extremely small, 5.3 Å for 235 aligned residue pairs. The probability alignment consisting of the most reliable 107 residue pairs even improves the r.m.s.d. to 2.6 Å.

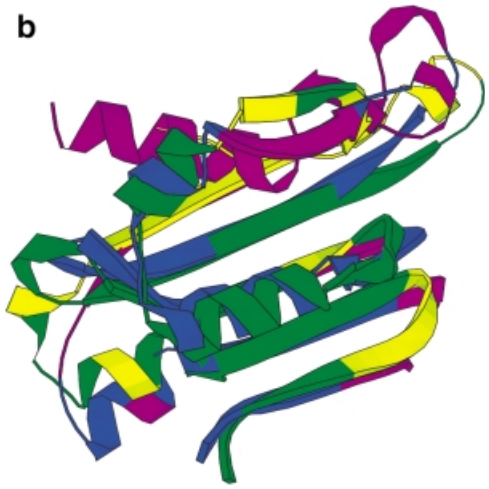

Fig. 8. (**a**) Sequence–structure alignments of glutathione reductase C-terminal domains of residues 364–478 (3GRS:364–478) and NADH peroxidase C-terminal domain of residues 322–447 (1NPX:322–447). Minimum energy score alignments and probability alignments are shown for both types of pairs, between the sequence of 3GRS:364–478 and the structure of 1NPX:322–447 and between the structure of 3GRS:364–478 and the sequence of 1NPX:322–447. The probability alignments are made by successively aligning site pairs in order of their alignment probabilities. The numbers below the sequences in these alignments represent probabilities with which those residue pairs are aligned; '5', for example, means that the probability is ⩾0.5 and <0.6. The question marks between sequences indicate that those site pairs do not correspond to site pairs with maximum alignment probabilities over all other sites and thus those alignments of residues are very uncertain. At the end of each alignment, the minimum energy score, the fraction of identical amino acid pairs, the number of aligned residues and the r.m.s.d. in superposition of those aligned residues are listed. For probability alignments, the r.m.s.d.s for residues aligned with probabilities ⩾0.5 are also listed. Characters 'a' and 'b' between the alignments show an assignment of α-helices and β-strands based on each protein structure; taken from the PDB files. Also, '#' between the alignments indicates site positions whose alignments are common in all alignments. The fraction of identical amino acid pairs and the r.m.s.d. for a conventional sequence alignment of the protein pair are listed in Table V. The energy scores per residue for native sequence and structure pairs are –1.15 for 3GRS:364–478 and –0.95 for 1NPX:322–447. (**b**) Structure of 3GRS (residues 364–478) (yellow and green) matched with the structure of 1NPX (residues 322–447) (magenta and blue) on the basis of the probability alignment of the 73 matched residues shown in (a) for the alignment of the sequence of 3GRS with the structure of 1NPX. These are residues all having a probability ⩾0.5 of being aligned by this method. The aligned residues are shown in green for 3GRS and in blue for 1NPX. Thus, it can be seen that the green and blue regions constitute a type of core of these structure fragments. The MOLSCRIPT program (Kraulis, 1991) was used to draw this figure; strands, turns and helices are assigned by the MOLAUTO program.

Effects of secondary structure potentials

To examine the effectiveness of secondary structure potentials on sequence–structure alignments, alignments are also calculated by including only tertiary structure energies, without secondary structure energies. In Figure 10, the fractions of identical amino acid pairs in alignments are compared between the two energy schemes, that is, with and without secondary structure energies. Alignments calculated with secondary structure energies tend to contain more identical amino acid pairs

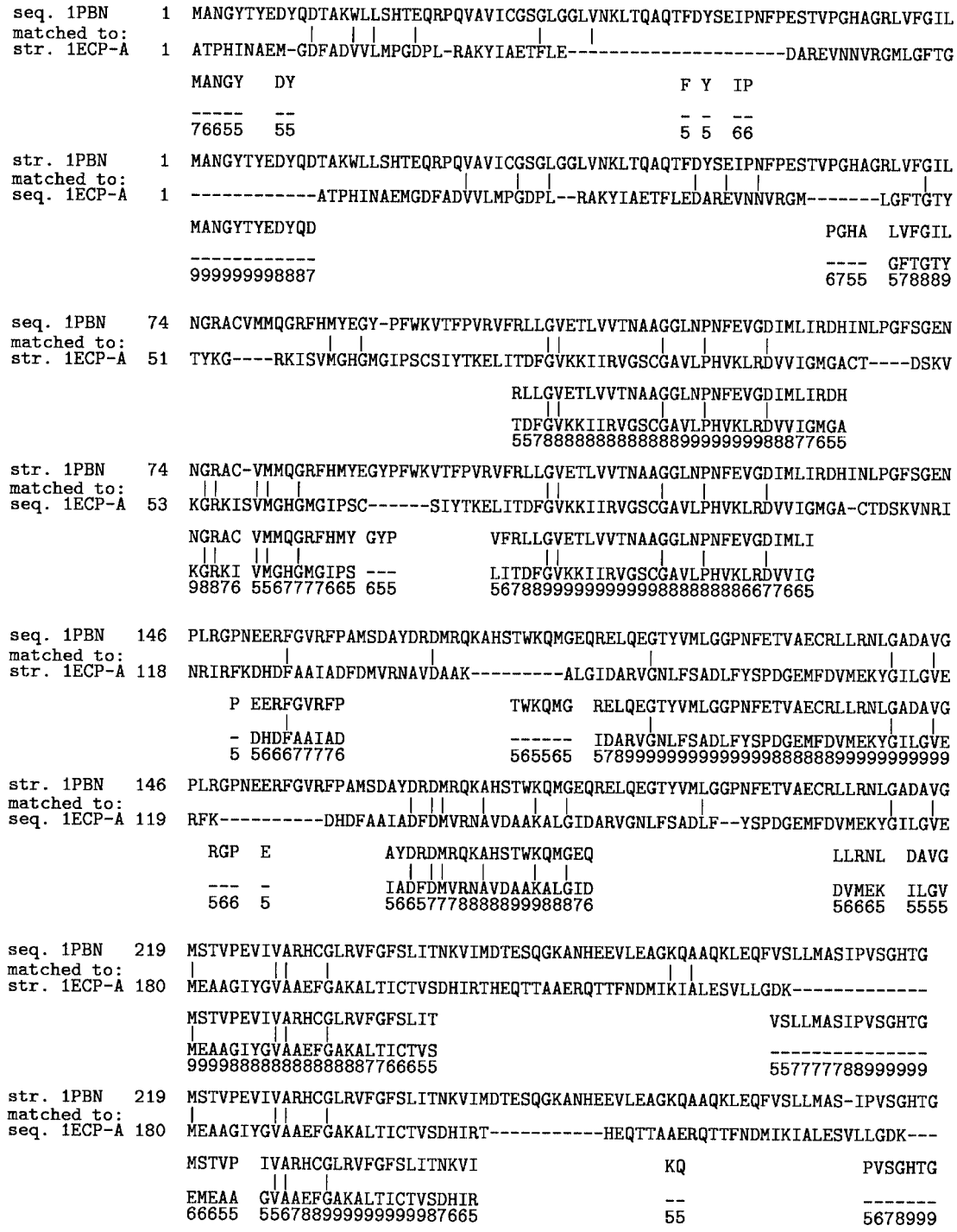

Fig. 9. Sequence–structure alignments of bovine purine nucleoside phosphorylase (1PBN) and purine nucleoside phosphorylase A chain from *E.coli* (1ECP-A). Minimum energy score alignments and probability alignments are shown for both types of pairs, between the sequence of 1PBN and the structure of 1ECP-A and between the structure of 1PBN and the sequence of 1ECP-A. In this figure, probability alignments are shown only for residues aligned with probabilities ⩾0.5. See the legend of Figure 8a for the numbers below the sequence and the question marks between sequences. The minimum energy scores, the r.m.s.d.s of aligned residues, the numbers of aligned residue pairs and the fractions of identical amino acid pairs for minimum energy score alignments are –6.5, 10.1 Å, 236 and 0.09 for 1PBN sequence versus 1ECP-A structure, and –14.7, 5.3 Å, 235 and 0.12 for 1PBN structure versus 1ECP-A sequence, respectively. The r.m.s.d.s of aligned residues and the numbers of aligned residue pairs for residues aligned with probabilities ⩾0.5 are 5.4 Å and 99 for 1PBN sequence versus 1ECP-A structure, and 2.6 Å and 107 for 1PBN structure versus 1ECP-A sequence, respectively. The energy scores per residue for native sequence and structure pairs are –1.13 for 1PBN and –1.01 for 1ECP-A.

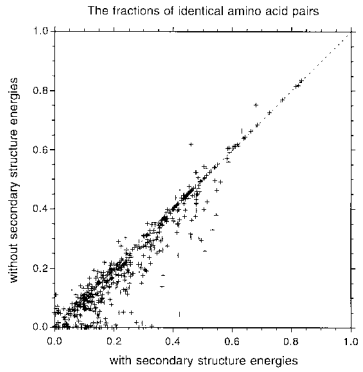

Fig. 10. Improvements in sequence–structure alignments obtained by including secondary structure potentials. The abscissas and ordinates show the fractions of identical amino acid pairs in sequence–structure alignments for 548 homologous protein pairs with and without secondary structure potentials, respectively. The dotted line shows a line with equal values for both axes. The values of parameters for alignments without secondary structure potentials are $1/\beta = 1.8$, $\mathcal{E}_0 = -0.57$, $w_0 = 4.17$, $w_1 = 0.63$, $w_2 = 0.36$ and $w_c = 45.57$, and those for terminal gaps are half of the corresponding parameters for middle gaps.

than those without secondary structure energies. This suggests that short-range energy potentials are useful to yield correct positions of residues in sequence–structure alignments. Also, as shown in Table IV, short-range energy potentials improve the capability for recognition of compatibility between sequences and structures. Such improvements in the recognition of sequence–structure compatibilities by secondary structure potentials were previously shown by threading sequences into structures without gaps (Miyazawa and Jernigan, 1999b). Figure 10 indicates that short-range energy potentials improve the recognition of sequence–structure compatibilities through yielding more correct positions of residues in sequence–structure alignments, even though long-range potentials work well principally for the recognition of overall folds. To obtain correct alignments, the short- and long-range potentials are complementary and both seem to be essential.

Discussion

An empirical potential, comprised of short-range secondary structure potentials and long-range contact energies together with repulsive packing potentials, has been tested to examine how well it can discriminate homologous structures from dissimilar folds for a given protein sequence, and also how well it can detect compatibilities between protein sequences and a given structure.

Miyazawa and Jernigan (1999c) reported that this same potential could discriminate native structures from non-native folds and also distinguish native sequences from non-native sequences, in which non-native pairs of sequences and structures are generated by threading in all possible ways, without gaps. In the present paper, significantly more non-native folds are generated by making sequence–structure alignments in which gaps in both the sequences and structures are permitted.

A scoring function to estimate the stabilities of protein sequence and structure pairs has been devised to assess compatibilities between sequences and structures. The compatibility between a sequence and a structure has been taken here to be equivalent to the stability for that pair of structure and sequence. On an energy scale of stability, folds can be compared with one another.

As discussed in our previous work (Miyazawa and Jernigan, 1999c), the following problems need to be solved. First, protein structures in multimeric states and monomeric states must be compared in order to judge which is more compatible with a given sequence. Since it is difficult to evaluate rigorously the stabilities of such folds in multimeric states, because of the entropy loss due to protein binding, we choose instead to approximate those stabilities. In order to overcome this problem, only the terms of conformational energy that depend on the amino acid sequence order have been included in the present energy function. In the case of contact energies, the collapse energy e_{rr} is subtracted from the contact energies (Miyazawa and Jernigan, 1996, 1999c). This modified energy scale was shown in Miyazawa and Jernigan (1996, 1999c) to provide a threading reference state for successfully discriminating native structures from non-native folds.

The second problem is more essential; the assessment of compatibilities between sequences and structures requires comparisons between different sequences for a given fold. Also, deletions and insertions in alignments must be considered in order to detect similar folds for a given sequence. As a result, the sequence dependences of the whole ensemble of protein conformations must be taken into account to measure stabilities of protein conformations. We take account of only dominant terms, i.e. native-like compact conformations in the summation of Boltzmann factors over all conformations, and then evaluate the logarithm of the partition function with the first and second terms in a high-temperature expansion. Finally, the zero energy state of the energy scoring function is adjusted for each sequence by representing conformational energies relative to a properly defined reference state, the conformational energy of a typical native structure with the same amino acid composition. For assessing the suitability of each type of residue for each structural position, the average conformational energy of each type of residue in the native structures has been chosen as a reference energy for that type of residue, relative to which conformational energies of folds are compared. In other words, sequence–structure alignments with a zero energy in this energy scale have conformational energies comparable to the native structure. It was shown in Miyazawa and Jernigan (1999c) that on the energy scale with this modification, native sequences had lower energy scores than all non-native sequences when the sequences were threaded into structures without gaps. Here it should be noted that in principle native structures ought to be the lowest energy folds for their sequences but native sequences need not be the most compatible with their native structures, even though this is highly probable; some proteins may be incompletely evolved toward the most compatible sequences.

As a result, this energy function with two types of modifications is expected to estimate properly the stabilities of protein structures with different sequences and also for different environments, i.e. monomeric and multimeric environments. The suitability of these modifications to the energy potentials for fold and sequence recognition is supported by the present results showing that this scoring function can recognize folds compatible with sequences, and inversely sequences with folds, and can generate mostly similar alignments for these two types of aligned sequence and structure pairs.

However, in order to allow deletions and insertions in sequence–structure alignments, additional parameters, corresponding to penalties for gaps, must be introduced into the scoring function. To obtain good alignments, it is important to use a proper gap scheme and to determine a set of appropriate

values for gap parameters. Lesk *et al*. (1986) pointed out that in globin sequences deletions and insertions are infrequently observed in the interiors of helical regions of proteins because of the importance of the stabilization for structures of the packing at helix–helix interfaces, and they introduced variable gap penalties between helical regions and inter-helical and loop regions. Barton and Sternberg (1987) also showed the superiority of their secondary structure dependent alignment method using various gap penalties. Fischel-Ghodsian *et al*. (1990) modified a dynamic programming method to include predicted secondary structure information. On the other hand, Kanaoka *et al*. (1989) assigned large gap penalties to the hydrophobic core. Ouzounis *et al*. (1993) demonstrated that the use of core weights considerably improves the detectability of remote homologues with sequence–structure alignments. In all of these analyses, sequence alignments were improved by the use of variable gap penalties. However, no structural information is available for most sequence alignments. Such a gap scheme is useful only for sequence–structure alignments.

Here, gap penalties are taken simply to be proportional to the number of contacts at each residue position in protein structures. The number of residue contacts is utilized as a simple measure of the packing density of residues. Thus, in densely packed regions in protein structures, insertions and deletions of residues rarely occur in alignments. If necessary, gap penalties could be set to depend on local secondary structures at each residue position. In this paper, we have not quantitatively examined how much the present gap scheme can improve sequence–structure alignments, although qualitative improvements are observed. It is difficult to determine an optimal set of values for gap parameters. It has also been shown that both sequence–structure alignments and conventional sequence alignments of homologous protein pairs have similar overall characteristics with respect to the proportions of deletions and identical residues (see Figure 1). However, it is not easy to obtain good alignments for proteins whose lengths are significantly different. Such alignments depend strongly on the gap parameters for termini. No penalty for terminal gaps may be better for aligning a single domain with a multi-domain protein for identifying a domain in multi-domain proteins, but it is not appropriate in other cases to ignore all terminal gaps.

Here, folds of multimeric proteins should always be evaluated in their multimeric states even against sequences of monomeric proteins. This is appropriate for searching for sequences compatible with a given structure. However, for searching for compatible folds with a given sequence, templates of protein folds should be evaluated in the monomeric state for monomeric sequences and in the multimeric state for multimeric sequences. Alternatively, protein folds could be evaluated in both the monomeric and multimeric states and the form with the lower energy chosen.

The present energy potential includes long-range, pairwise interaction potentials, so that the exact solution of a minimum energy score alignment cannot be calculated by the dynamic programming method. It is another technical problem in fold recognition to obtain minimal energy-score alignments with such a long-range potential.

Here, pairwise contact energies have been evaluated in a mean field approximation on the basis of probabilities of site pairs being aligned. To obtain self-consistent values of alignment probabilities of site pairs, an iterative method is used for pairwise potential extraction. In most cases, iterations converge rapidly. This approximation becomes rigorous in the low-temperature limit, but it is more useful at higher temperature where an ensemble of alignments becomes significant rather than a minimum energy alignment. In addition to the most probable alignment, i.e. the minimum energy alignment, an alignment has also been made by successively aligning site pairs in order of their alignment probabilities. This alignment method based on alignment probabilities of site pairs is consistent with the mean field approximation for pairwise contact energies to be evaluated on the basis of those probabilities. Alignments made by this probability alignment method coincide with the most probable alignments in a low-temperature limit. This method also provides information about how reliable each aligned site pair is. Figure 2 indicates that alignments consisting of residues aligned with high probabilities can improve significantly the r.m.s.d.s in superposition of two proteins. This feature is particularly desirable for aligning distantly related sequences and structures. Also, it is noteworthy that reliably aligned residue pairs between the sequence of 3GRS and the structure of 1NPX constitute a type of core of these structure fragments (see Figure 8b).

It has been clearly demonstrated that the present scoring function including the present modifications in energy scale and parameters for gap penalties can properly evaluate compatibilities between sequences and structures (see Figures 1–4), and therefore can be used both for searching for compatible folds with a given sequence and likewise for searching for compatible sequences with a given fold (see Figure 3). Figure 5 shows that this method of sequence–structure alignment complements conventional sequence alignment in detecting compatible proteins. As shown in Table V, it is further useful to find a significant number of new protein pairs in which structures are similar but in which sequences were too different for the conventional sequence alignments to detect their structural similarities.

References

Barton,G.J. and Sternberg,M.J.E. (1987) *Protein Eng*., **1**, 89–94.

Bowie,J.U., Lüthy,R. and Eisenberg,D. (1991) *Science*, **253**, 164–170.

Bryant,S.H. and Lawrence,C.E. (1993) *Proteins* **16**, 92–112.

Crippen,G.M. (1991) *Biochemistry*, **30**, 4232–4237.

Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) In Dayhoff,M.O. (ed.). *Atlas of Protein Sequence and Structure 1978*, vol. 5, Suppl. 3. National Biomedical Research Foundation, Washington, DC, pp. 345–352.

Feller,W. (1968) *An Introduction to Probability Theory and its Applications*, vol. I. Wiley, New York.

Finkelstein,A.V. and Reva,B.A. (1991) *Nature*, **351**, 497–499.

Fischel-Ghodsian,F., Mathiowitz,G. and Smith,T.F. (1990) *Protein Eng*., **3**, 577–581.

Fitch,W.M. and Smith,T.F. (1983) *Proc. Natl Acad. Sci. USA*, **80**, 1382–1386.

Go,M. and Miyazawa,S. (1980) *Int. J. Pept. Protein Res*., **15**, 211–224.

Godzik,A., Kolinski,A. and Skolnick,J. (1992) *J. Mol. Biol*., **227**, 227–238.

Gotoh,O. (1990) *Bull. Math. Biol*., **52**, 359–373.

Hendlich,M., Lackner,P., Weitckus,S., Floechner,H., Froschauer,R., Gottsbachner,K., Casari,G. and Sippl,M.J. (1990) *J. Mol. Biol*., **216**, 167–180.

Henikoff,S. and Henikoff,J.G. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Huang,E.S., Subbiah,S. and Levitt,M. (1995) *J. Mol. Biol*., **252**, 709–720.

Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) *Nature*, **358**, 86–89.

Jones,D. and Thornton,J. (1993) *J Comput.-Aided Mol. Des*., **7**, 439–456.

Kanaoka,M., Kishimoto,F., Ueki,Y. and Umeyama,H. (1989) *Protein Eng*., **2**, 347–351.

Karlin,S. and Altschul,S.F. (1990) *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.

Kocher,J.-P.A., Rooman,M.J. and Wodak,S.J. (1994) *J. Mol. Biol*., **235**, 1598–1613.

Kraulis,P.J. (1991) *J. Appl. Crystallogr*., **24**, 946–950.

Lathrop,R.H. and Smith,T.F. (1996) *J. Mol. Biol*., **255**, 641–665.

Lesk,A.M., Levitt,M. and Chothia,C. (1986) *Protein Eng*., **1**, 77–78.
Lüthy,R., Bowie,J.U. and Eisenberg,D. (1992) *Nature*, **356**, 83–85.
Maiorov,V.N. and Crippen,G.M. (1992) *J. Mol. Biol*., **227**, 876–888.
Matsuo,Y., Nakamura,H. and Nishikawa,K. (1995) *J. Biochem*., **118**, 137–148.
Matsuo,Y. and Nishikawa,K. (1994) *Protein Sci*., **3**, 2055–2063.
Mirny,L.A. and Shakhnovich,E.I. (1996) *J. Mol. Biol*., **264**, 1164–1179.
Miyazawa,S. (1995) *Protein Eng*., **8**, 999–1009.
Miyazawa,S. and Jernigan,R.L. (1985) *Macromolecules*, **18**, 534–552.
Miyazawa,S. and Jernigan,R.L. (1996) *J. Mol. Biol*., **256**, 632–644.
Miyazawa,S. and Jernigan,R.L. (1999a) *Proteins* **34**, 49–68.
Miyazawa,S. and Jernigan,R.L. (1999b) *Proteins* **36**, 347–356.
Miyazawa,S. and Jernigan,R.L. (1999c) *Proteins* **36**, 357–369.
Moews,P.C. and Kretsinger,R.H. (1975) *J. Mol. Biol*., **91**, 201–228.
Munson,P.J. and Singh,R.K. (1997) *Protein Sci*., **6**, 1467–1481.
Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol*., **247**, 536–540.
Needleman,S.B. and Wunsch,C.B. (1970) *J. Mol. Biol*., **48**, 443–453.
Nishikawa,K. and Matsuo,Y. (1993) *Protein Eng*., **6**, 811–820.
Ouzounis,C., Sander,C., Scharf,M. and Schneider,R. (1993) *J. Mol. Biol*., **232**, 805–825.
Park,B. and Levitt,M. (1996) *J. Mol. Biol*., **258**, 367–392.
Park,B.H., Huang,E.S. and Levitt,M. (1997) *J. Mol. Biol*., **266,** 831–846.
Samudrala,R. and Moult,J. (1998) *J. Mol. Biol*., **275**, 895–916.
Sippl,M.J. (1990) *J. Mol. Biol*., **213**, 859–883.
Sippl,M.J. (1993) *Proteins*, **17**, 355–362.
Sippl,M.J. and Weitckus,S. (1992) *Proteins*, **13**, 258–271.
Taylor,W.R. (1997) *J. Mol. Biol*., **269**, 902–943.
Taylor,W.R. and Orengo,C.A. (1989) *J. Mol. Biol*., **208**, 1–22.
Thomas,P.D. and Dill,K.A. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 11628–11633.
Vendruscolo,M. and Domany,M. (1998) *Folding Des*., **3**, 329–336.
Vingron,M. and Waterman,M.S. (1994) *J. Mol. Biol*., **235**, 1–12.

Received October 4, 1999; revised March 6, 2000; accepted April 12, 2000